Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model

Supplementary Material

In the supplementary document, we will introduce the following contents: 1) details of **TPS transformation** (Sec. A); 2) more details of our proposed **framework** (Sec. B), including the motion decoupling module (Sec. B.1), the latent motion diffusion model (Sec. B.2), the refinement network (Sec. B.3), the optimal motion selection module (Sec. B.4), and other implementation details (Sec. B.5); 3) the selection of **objective metrics** (Sec. C); 4) more details and analysis of **comparison to existing methods** (Sec. D); 5) results and analysis of the **ablation study** (Sec. E); 6) capability of generating **long gesture videos** (Sec. F); 7) **user study** details (Sec. G); 8) analysis of the **robustness and effectiveness of objective metrics** (Sec. H); 9) **generalization ability** analysis (Sec. I); 10) **time and resource consumption** (Sec. J); 11) **limitations and future work** (Sec. K); 12) **dataset license** (Sec. L). Since more mathematical expressions are included, we choose a single-column format in this supplementary document instead of two-column for readability. All demos, code, and more resources can be found at `https://github.com/thuhcsi/S2G-MDDiffusion`.

A. Details of TPS Transformation

In the main paper, we employ TPS transformation [3] to establish pixel-level optical flow relying solely on sparse keypoint pairs from driving and reference images, thereby achieving precise control over the motion of human body regions. This is the foundation of our approach to decoupling motion while retaining crucial appearance information. Here we give a more detailed explanation of TPS transformation.

TPS transformation is a type of image warping algorithm. It takes as input corresponding N pairs of keypoints $(p_i^{\mathbf{D}}, p_i^{\mathbf{S}}), i = 1, 2, \ldots, N$ (referred to as control points) from a driving image $\mathbf{D}$ and a source image $\mathbf{S}$, and outputs a pixel coordinate mapping $\mathcal{T}_{tps}(\cdot)$ from $\mathbf{D}$ to $\mathbf{S}$ (referred to as backward optical flow). This process is grounded in the foundational assumption that the 2D warping can be emulated through a thin plate deformation model. TPS transformation seeks to minimize the energy function necessary to bend the thin plate, all while ensuring that the deformation accurately aligns with the control points, and the mathematical formulation is as follows:

$$\begin{aligned} \min \iint_{\mathbb{R}^2} & \left(\left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 \mathcal{T}_{tps}}{\partial y^2} \right)^2 \right) dxdy, \\ \text{s.t.} \quad & \mathcal{T}_{tps}(p_i^{\mathbf{D}}) = p_i^{\mathbf{S}}, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{A1}$$

where $p_i^{\mathbf{D}}$ and $p_i^{\mathbf{S}}$ denotes the i^{th} keypoints paired in $\mathbf{D}$ and $\mathbf{S}$. According to [3], it can be proven that TPS interpolating function is a solution to Eq. (A1):

$$\mathcal{T}_{tps}(p) = A \begin{bmatrix} p \\ 1 \end{bmatrix} + \sum_{i=1}^{N} w_i U\left(\left\| p_i^{\mathbf{D}} - p \right\|_2 \right), \tag{A2}$$

where $p = (x, y)^\top$ is the origin coordinate in $\mathbf{D}$, and $p_i^{\mathbf{D}}$ is the i^{th} keypoint in $\mathbf{D}$. $U(r) = r^2 \log r^2$ is a radial basis function. Actually, $U(r)$ is the fundamental solution of the biharmonic equation [5] that satisfies

$$\Delta^2 U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 U \propto \delta_{(0,0)}, \tag{A3}$$

where the generalized function $\delta_{(0,0)}$ is defined as

$$\delta_{(0,0)} = \begin{cases} \infty, & \text{if } (x, y) = (0, 0) \\ 0, & \text{otherwise} \end{cases}, \quad \text{and} \iint_{\mathbb{R}^2} \delta_{(0,0)}(x, y)\, dxdy = 1, \tag{A4}$$

which means that $\delta_{(0,0)}$ is zero everywhere except at the origin while having an integral equal to 1.

We use $p_i^{\mathbf{X}} = (x_i^{\mathbf{X}}, y_i^{\mathbf{X}})^\top$ to denote the i^{th} keypoint in image $\mathbf{X}$ (*i.e.* $\mathbf{D}$ or $\mathbf{S}$), and denote:

$$r_{ij} = \left\| p_i^{\mathbf{D}} - p_j^{\mathbf{D}} \right\|, \quad i, j = 1, 2, \ldots, N,$$

$$K=\left[\begin{array}{cccc} 0 & U\left(r_{12}\right) & \cdots & U\left(r_{1N}\right) \\ U\left(r_{21}\right) & 0 & \cdots & U\left(r_{2N}\right) \\ \vdots & \vdots & \ddots & \vdots \\ U\left(r_{N1}\right) & U\left(r_{N2}\right) & \cdots & 0 \end{array}\right], \quad P=\left[\begin{array}{ccc} 1 & x_1^{\mathbf{D}} & y_1^{\mathbf{D}} \\ 1 & x_2^{\mathbf{D}} & y_2^{\mathbf{D}} \\ \vdots & \vdots & \vdots \\ 1 & x_N^{\mathbf{D}} & y_N^{\mathbf{D}} \end{array}\right],$$

$$L=\left[\begin{array}{cc} K & P \\ P^T & 0 \end{array}\right], \quad Y=\left[\begin{array}{cccccccc} x_1^{\mathbf{S}} & x_2^{\mathbf{S}} & \cdots & x_N^{\mathbf{S}} & 0 & 0 & 0 \\ y_1^{\mathbf{S}} & y_2^{\mathbf{S}} & \cdots & y_N^{\mathbf{S}} & 0 & 0 & 0 \end{array}\right]^{\top}.$$

Then we can solve the affine parameters $A \in \mathcal{R}^{2\times 3}$ and TPS parameters $w_i \in \mathcal{R}^{2\times 1}$ as:

$$\left[w_1, w_2, \cdots, w_N, A\right]^{\top} = L^{-1}Y. \tag{A5}$$

In fact, in Eq. (A2), the first term $A\left[\begin{array}{c} p \\ 1 \end{array}\right]$ is an affine transformation for the alignment in the linear space of paired control points $(p_i^{\mathbf{D}}, p_i^{\mathbf{S}})$. The second term $\sum_{i=1}^{N} w_i U\left(\left\|p_i^{\mathbf{D}} - p\right\|_2\right)$ introduces non-linear distortions for elevating or lowering the thin plate. With both the linear and nonlinear transformations, TPS transformation allows for precise deformation which is important to describe the motion without discarding crucial appearance information in our framework.

B. More Details of Our Proposed Framework

B.1. Motion Decoupling Module

Training losses. The motion decoupling module is trained end-to-end in an unsupervised manner. From previous works [29, 30, 42], we use a pretrained VGG-19 network [31] to calculate the perceptual construction loss in different resolutions as the main driving loss:

$$\mathcal{L}_{per} = \sum_j \sum_i \left| \mathrm{VGG19}_i\left(\mathrm{DS}_j(\mathbf{D})\right) - \mathrm{VGG19}_i(\mathrm{DS}_j(\hat{\mathbf{D}}))\right|, \tag{B6}$$

where $\mathrm{VGG19}_i$ means the i^{th} layer of the VGG-19 network, while DS_j represents j downsampling operations. Also, equivariance loss is used to enhance the stability of the keypoint predictor as:

$$\mathcal{L}_{eq} = \left| E_{kp}(\widetilde{\mathcal{A}}(\mathbf{S})) - \widetilde{\mathcal{A}}\left(E_{kp}(\mathbf{S})\right)\right|, \tag{B7}$$

where E_{kp} is the keypoint predictor, and $\widetilde{\mathcal{A}}$ is a random geometric transformation operator.

In addition, as introduced in [42], we also encode $\mathbf{D}$ into feature maps with the encoder of the image synthesis network, compared with warped reference feature maps to calculate the warping loss:

$$\mathcal{L}_{warp} = \sum_i \left| \widetilde{\mathcal{T}}^{-1}\left(E_i(\mathbf{S})\right) - E_i(\mathbf{D})\right|, \tag{B8}$$

where E_i is the i^{th} layer of the encoder of the image synthesis network, and $\widetilde{\mathcal{T}}^{-1}$ denotes the inverse function of the estimated optical flow, *i.e*. the forward optical flow from $\mathbf{R}$ to $\mathbf{D}$.

The final loss is the sum of the above terms:

$$\mathcal{L}_{tps} = \mathcal{L}_{per} + \mathcal{L}_{eq} + \mathcal{L}_{warp}. \tag{B9}$$

B.2. Latent Motion Diffusion Model

Framework. The framework of our latent motion diffusion model is based on DDPM [8], where diffusion is defined as a Markov noising process. $\boldsymbol{x}_0 \sim p(\boldsymbol{x})$ is sampled from the real data distribution (*i.e*. $\boldsymbol{x}_0$ is a sequence of latent motion features drawn from a real gesture video). Given constant hyper-parameters $\alpha_t \in (0,1)$ decreasing with t, the forward diffusion process is to add Gaussian noise to the sample:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \left(1-\alpha_t\right)\mathbf{I}\right). \tag{B10}$$

When the maximum time step T is sufficiently large and α_t is small enough, we can use standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to approximate $\boldsymbol{x}_T$. This indicates that it is possible to estimated real posterior $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ following the reverse denoising process:

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)), \tag{B11}$$

where $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$ mean estimating the mean and covariance via a neural network with learnable parameters θ. From DDPM [8], the network predicts the noise $\epsilon_\theta(\boldsymbol{x}_t, t)$ and thus we can use $\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t, t)\right)$ added by randomly sampled noise to estimate $\boldsymbol{x}_{t-1}$. In our context, we take speech audio and the seed motion feature of the reference frame as conditions c, and aim to model the conditional distribution $p_\theta(\boldsymbol{x}_0|c)$ by gradually removing the noise. Following [24], we predict $\boldsymbol{x}_0$ itself instead of noise ϵ. The neural network of the diffusion network can be represented as $\hat{\boldsymbol{x}}_0 = \mathcal{G}(\boldsymbol{x}_t, t, c)$.

Training losses. We follow [8] to use *simple* objective as the first term of losses:

$$\mathcal{L}_{simple} = E_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}|c), t \sim [1,T]}\left[\|\boldsymbol{x}_0 - \mathcal{G}(\boldsymbol{x}_t, t, c)\|_2^2\right]. \tag{B12}$$

Besides, as mentioned in the main paper, we use the velocity loss and the acceleration loss to constrain the physical attributes of the motion features that describe the trajectories of the keypoint movements. Velocity and acceleration are respectively defined as the first and second-order time derivatives of the keypoint positions, and here, differential methods are employed to represent derivatives [32, 33]:

$$\mathcal{L}_{vel} = \frac{1}{M-1}\sum_{m=1}^{M-1}\left\|\left(\boldsymbol{x}_0^{(m+1)} - \boldsymbol{x}_0^{(m)}\right) - \left(\hat{\boldsymbol{x}}_0^{(m+1)} - \hat{\boldsymbol{x}}_0^{(m)}\right)\right\|_2^2, \tag{B13}$$

$$\mathcal{L}_{acc} = \frac{1}{M-2}\sum_{m=1}^{M-2}\left\|\left[\left(\boldsymbol{x}_0^{(m+2)} - \boldsymbol{x}_0^{(m+1)}\right) - \left(\boldsymbol{x}_0^{(m+1)} - \boldsymbol{x}_0^{(m)}\right)\right] - \left[\left(\hat{\boldsymbol{x}}_0^{(m+2)} - \hat{\boldsymbol{x}}_0^{(m+1)}\right) - \left(\hat{\boldsymbol{x}}_0^{(m+1)} - \hat{\boldsymbol{x}}_0^{(m)}\right)\right]\right\|_2^2. \tag{B14}$$

The final training loss is as follows:

$$\mathcal{L}_{diff} = \mathcal{L}_{simple} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{acc}\mathcal{L}_{acc}. \tag{B15}$$

Guidance. Following [33], we train our diffusion model with classifier-free guidance. In training, we randomly mask the speech audio with a certain probability of 25%, *i.e.* replacing the condition $c = \{\boldsymbol{a}, x_0^{(0)}\}$ with $c_\varnothing = \{\varnothing, x_0^{(0)}\}$. Then, we can strike a balance between diversity and fidelity by weighting the two results with γ:

$$\hat{\boldsymbol{x}}_0 = \gamma\mathcal{G}(\boldsymbol{x}_t, t, c) + (1-\gamma)\mathcal{G}(\boldsymbol{x}_t, t, c_\varnothing), \tag{B16}$$

where we can use $\gamma > 1$ for extrapolating to enhance the speech condition.

B.3. Refinement Network

Architecture details. Inspired by [23], we use a Unet-like [26] architecture to restore missing details of synthesized image frames. In specific, we use eight "`convolution - LeakyReLU - batch norm`" downsampling blocks and eight "`upsample - convolution - LeakyReLU - batch norm`" upsampling blocks with long skip connections, which prevent the information loss during downsampling while maintaining a large receptive field. Additionally, we insert two residual blocks [40] into the final two layers respectively, whose shallow architecture leads to a small receptive field and processes the feature maps in a sliding window manner. Simultaneously possessing large and small receptive fields enables the refinement network to capture both global and local information, thus better recovering missing details. Also, to ensure authenticity, we employ a patch-based discriminator [23] trained with GAN discriminator loss $\mathcal{L}_D$ for adversarial training. Both the ground truth and refined image are converted into feature maps, with each element being discriminated as real or fake.

Training losses. Firstly, we train the refinement network with the common L1 reconstruction loss. Note that, as mentioned in the main paper, we utilize MobileSAM [41] to segment hands and the face to get the masks, and assign larger weights to both hands, face, and occluded areas using the masks in L1 reconstruction loss:

$$\mathcal{L}_{rec} = \mathcal{L}_{valid} + \lambda_{occ}\mathcal{L}_{occ} + \lambda_{hand}\mathcal{L}_{hand} + \lambda_{face}\mathcal{L}_{face}, \tag{B17}$$

where we use the complement of the occlusion masks from the optical flow predictor to compute $\mathcal{L}_{valid}$.

Then similar to [17, 23, 35], VGG-16 [31] is used to compute the perceptual loss and style loss in the feature space as:

$$\mathcal{L}_{per} = \sum_i \left| \text{VGG16}_i(\mathbf{D}) - \text{VGG16}_i(\hat{\mathbf{D}}_{ref}) \right|, \tag{B18}$$

$$\mathcal{L}_{style} = \sum_i \left| \text{VGG16}_i(\mathbf{D}) \cdot [\text{VGG16}_i(\mathbf{D})]^\top - \text{VGG16}_i(\hat{\mathbf{D}}_{ref}) \cdot [\text{VGG16}_i(\hat{\mathbf{D}}_{ref})]^\top \right|, \tag{B19}$$

where $\hat{\mathbf{D}}_{ref}$ and $\mathbf{D}$ represent the refined image frame and the real image frame respectively. VGG16$_i$ means the i^{th} layer of the VGG-16 network, and we select $i = 5, 10, 17$ in this work. In addition, following [17, 23], the total variation (TV) loss is used as:

$$\mathcal{L}_{tv} = \sum_i \sum_j \left(\left| \hat{\mathbf{D}}_{ref}^{i+1,j} - \hat{\mathbf{D}}_{ref}^{i,j} \right| + \left| \hat{\mathbf{D}}_{ref}^{i,j+1} - \hat{\mathbf{D}}_{ref}^{i,j} \right| \right), \tag{B20}$$

where $\hat{\mathbf{D}}_{ref}^{i,j}$ denotes the (i, j) pixel of the refined image frame.

The final loss is the weighted sum of the above terms, along with GAN generator loss $\mathcal{L}_G$:

$$\mathcal{L}_{ref} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{style}\mathcal{L}_{style} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_G\mathcal{L}_G. \tag{B21}$$

B.4. Optimal Motion Selection Module

We employ a segment-wise generation approach to generate motion feature sequences of arbitrary length. Inspired by [14], starting from the second segment, leveraging the diversity generation capability of diffusion, we generate P candidates for each segment conditioned on the current audio and the end frame of the preceding segment. The scores are computed using the last five frames of the preceding segment and the first five frames of the candidate.

Specifically, by reorganizing the motion features back into keypoint positions, we calculate two scores: 1) Position coherency score calculates the **L1 distance** between the mean positions of the preceding segment and all candidates over five frames. 2) Velocity consistency score calculates the **angle of velocity directions** in average between the preceding and candidate segments over five frames, where velocity is computed through the differential of position. These two scores are summed to obtain the final score. A lower final score indicates fewer abrupt changes in position and velocity direction between two segments, thereby reducing flickers and jitters. So the candidate segment with the lowest score is chosen to extend the motion feature sequence. The frames at the transition points are eventually filled using cubic spline interpolation.

B.5. Other Implementation Details

We train our overall framework on four speakers jointly in three stages. 1) For the motion decoupling module: The number of TPS transformations K is set to 20, each with $N = 5$ paired keypoints. We select ResNet18 [7] as the keypoint predictor for its simplicity and modify its output dimension to $20 \times 5 \times 2$ to match the number and dimension of keypoints. Following [42], the optical flow predictor and the image synthesis network are 2D-convolution-based and produce 64×64 weight maps to generate optical flow and four occlusion masks of different resolutions (32, 64, 128, 256) to synthesize image frames. We conduct training using Adam optimizer [10] with learning rate of 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$. 2) For the latent motion diffusion model: Keypoints are gathered and unfolded into the motion feature $x \in \mathcal{R}^{200}$ for each frame. Motion features and audios are clipped to $M = 80$ frames (3.2s) with stride 10 (0.4s) for training. The 35-dimension hand-crafted audio features include MFCC, constant-Q chromagram, tempogram, on-set strength and on-set beat, which are concatenated with 1024-dimension WavLM features to form $\boldsymbol{a} \in \mathcal{R}^{1059}$. For Eq. (B15), we set $\lambda_{vel} = \lambda_{acc} = 1$ and use Adan optimizer [36] with learning rate of 2×10^{-4} and 0.02 weight decay for 3,000 epochs training. The maximum sampling step T is 50. 3) For the refinement network: We set $\lambda_{occ} = 3$, $\lambda_{hand} = \lambda_{face} = 5$ in Eq. (B17). Following the hyper-parameter search results in [17], we set $\lambda_{per} = 0.05$, $\lambda_{style} = 120$, $\lambda_{tv} = 0.1$, and $\lambda_{GAN} = 0.1$ in Eq. (B21). Adam optimizer [10] with learning rate of 2×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$ is used for the refinement generator and learning rate of 4×10^{-5} for the discriminator. The whole framework is trained on 6 NVIDIA A10 GPUs for 5 days. In inference, γ in Eq. (B16) is set to 2 for extrapolating to augment the speech condition. Candidate number P is set to 5 for the balance between quality and inference time.

C. Selection of Objective Metrics

As a relatively unexplored task, co-speech gesture video generation lacks effective means of objective evaluation. Pioneering work ANGIE [18] simplifies the evaluation process by degrading their generation framework to 2D human skeletons

Table D1. **Subjective evaluation results on test set with two generation schemes for MM-Diffusion.** Bold indicates the best and underline indicates the second. Results of MOS are presented with 95% confidence intervals. Only the favorable results of MM-Diffusion-C are reported in the main paper.

Name	Subjective evaluation			
	Realness ↑	Diversity ↑	Synchrony ↑	Overall quality ↑
Ground Truth (GT)	4.76±0.05	4.70±0.06	4.77±0.05	4.73±0.06
ANGIE	2.07±0.08	2.53±0.08	2.19±0.08	2.00±0.07
MM-Diffusion-D	1.63±0.09	1.98±0.09	1.54±0.08	1.46±0.08
MM-Diffusion-C	1.77±0.08	2.02±0.09	1.69±0.08	1.47±0.07
Ours	**3.79±0.08**	**3.91±0.07**	**3.90±0.08**	**3.77±0.07**

before leveraging the objective metrics common in skeleton generation, which, however, only assesses the performance of the generation module in structural skeletons without considering the effectiveness of the entire framework for gesture video generation. [43] employs metrics such as LPIPS popular in image evaluation and MOVIE for video evaluation to assess gesture reenactment. However, these general visual metrics only operate in the pixel or pixel-derived feature space, neglecting the crucial body movements in gesture videos. Therefore, we propose to use both motion and video-related metrics to evaluate gesture videos. Specifically, we use **Fréchet Gesture Distance** (**FGD**) [38], **Diversity** (**Div.**) [19], and **Beat Alignment Score** (**BAS**) [13] to evaluate the motion quality, and use **Fréchet Video Distance** (**FVD**) [34] to evaluate the video quality.

Details of motion-related metrics. We first extract 2D human poses with off-the-shelf pose estimator MMPose [28]. Extracting poses after generating gesture videos avoids the degradation of our original generation framework, allowing for effective measurements of the gesture motion quality in the videos. For the feasibility of calculating metrics, we performed normalization on raw poses: 1) We preserve 13 keypoints for the upper body and 21 keypoints for each hand, 55 keypoints in total [15, 28]. 2) We align the wrist points from body detection with those from hand detection. 3) For frames where the body is not detected, all keypoints are defined as centered at (128, 128). 4) For frames where hands are not detected, 21×2 hand keypoints are assigned to the corresponding body wrist points.

Then, BAS can be directly computed using the audio and the normalized poses. For FGD and Diversity metrics, we follow [22] to train an auto-encoder on pose sequences from PATS train set to encode poses into a feature space. During training, pose sequences are clipped to 80 frames without overlapping. Each clip is then encoded into a 32-dimension feature. For FGD, we compute the Fréchet Distance between features of generated videos and all real videos, including both train set and test set. For Diversity, we calculate the average Euclidean distance of generated videos in the feature space following [19].

D. Comparison to Existing Methods

As stated in the main paper, we compare our method with ANGIE [18] and MM-Diffusion [27]. For both our method and ANGIE, we use the audio and the initial frame image from PATS test set as inputs to generate corresponding 25fps gesture videos with a resolution of 256×256. Given that MM-Diffusion is trained solely conditioned on audio segments to generate 1.6s video segments of 10fps, we implement it with two generation schemes: 1) directly sampling long noise to generate videos of corresponding audio length (MM-Diffusion-D) and, 2) generating 1.6s segments for concatenation (MM-Diffusion-C). For both schemes, the generated gesture videos are resampled to 25fps. Additionally, considering that our method and MM-Diffusion-C generate fixed-length sub-clips (3.2s and 1.6s respectively) to form the full videos, both ground truth and generated videos are cropped to multiples of 3.2s for fair comparison.

User study results, including both of the two generation schemes of MM-Diffusion, are presented in Tab. D1. Due to space constraints, only the favorable results (MM-Diffusion-C) are reported in the main paper as "MM-Diffusion". It is important to note that MM-Diffusion does not use the initial frame image as a condition, thus lacking control over the appearance of the speaker in the generated videos, resulting in inconsistent speakers between concatenated segments. So, in the user study, participants are instructed to evaluate the videos generated by MM-Diffusion-C only within each 1.6s segment, neglecting the overall quality of the full-length video. This, in fact, is a lenient evaluation for disregarding the inherent limitation of MM-Diffusion in generating consistently long videos. Nonetheless, the experimental results still demonstrate the superiority of our method over MM-Diffusion in all dimensions. Despite some setting differences, this concessive evaluation is sufficient to prove that our method surpasses MM-Diffusion when generating short segments in gesture-specific scenarios, not to mention the capability of our method to generate consistent long gesture videos.

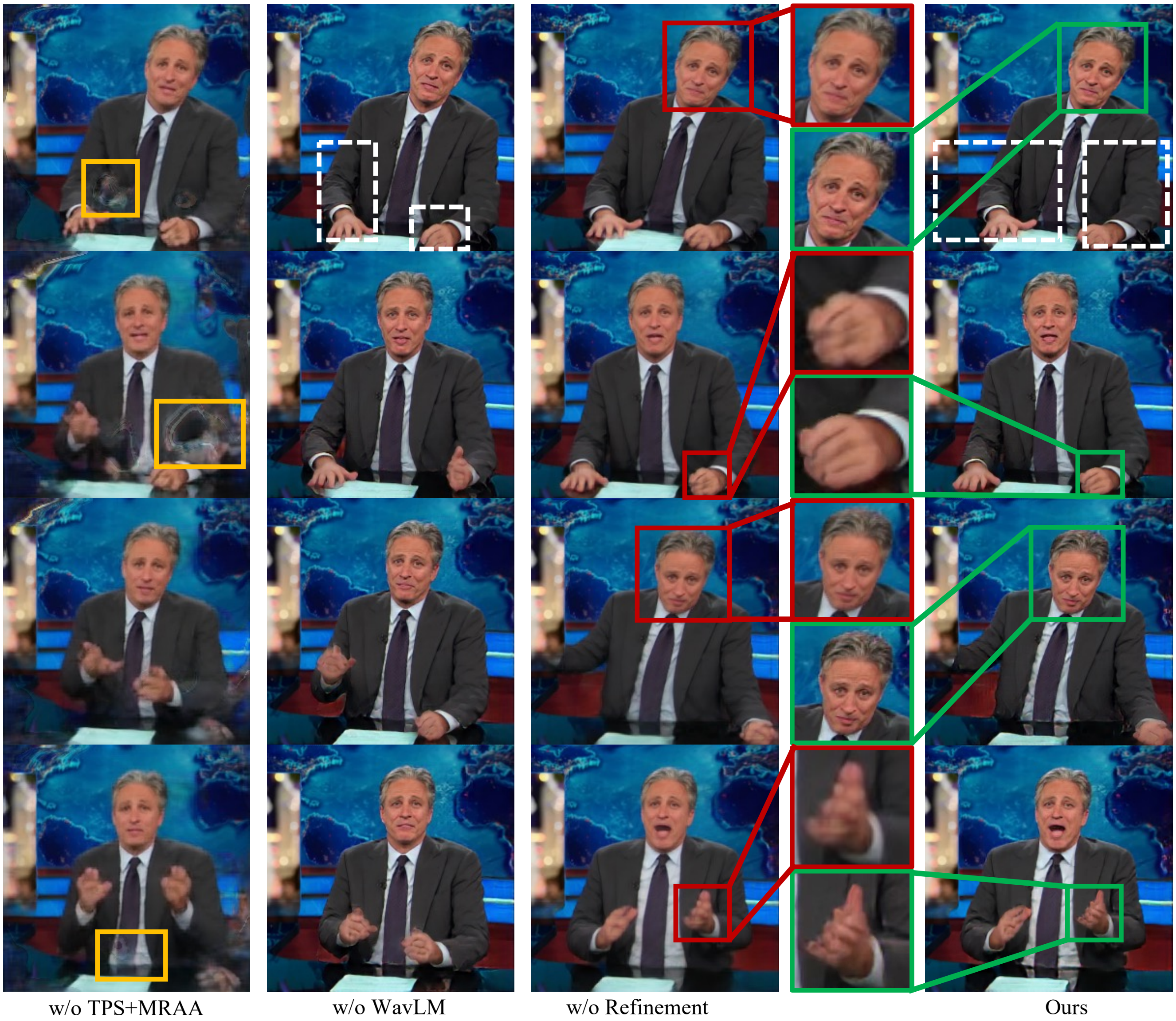


Figure E1. **Visualization results of the ablation study.** Replacing TPS with MRAA leads to ghost effects (yellow boxes). WavLM brings greater amplitude of hand motion (dashed boxes) given an impassioned speech. Refinement restores the details especially in hands and the face (red and green boxes).

Constrained by computational resources and referring to the result of our user study in Tab. D1, only the favorable MM-Diffusion-C is used to generate 480 test videos for objective evaluation and reported as "MM-Diffusion" in the main paper.

E. Ablation Study

Visualization results of the ablation study are shown in Fig. E1, where an impassioned speech is given as the condition. From the first column, we observe that the generated videos exhibit severe ghost effects (labeled by yellow boxes) when we replace the TPS-based motion features with MRAA [30]. We will give an explanation in the following part. According to [30], MRAA is a PCA-based affine transformation that represents motion features as the mean μ and the covariance Σ of the probability distribution of body regions. While it is appropriate to infer μ as the region translation from speech, the interaction between speech and the region shape represented by Σ is quite unclear. Unlike ANGIE [18] which uses a cross-condition GPT to connect Σ with μ and speech, our diffusion model emphasizes the interactions between speech and motion

Table F2. **Results of generating long gesture videos.** Bold indicates the best and underline indicates the second.

Name	Effective duration ↑
Ground Truth (GT)	27.8s
ANGIE	4.1s
LN Samp.	3.5s
Concat.	15.9s
Ours	**21.0s**

features, with less focus on relating Σ to μ. Thus the prediction of Σ is unstable. Although we impose constraints on Σ to be symmetric positive definite using Cholesky decomposition as mentioned in [18] for valid gestures, it still tends to output near-singular matrices, resulting severe errors in heatmaps for the estimation of the optical flow and occlusion masks. This, in turn, causes undesirable visual effects.

The second column shows the results of removing WavLM [4] features with only hand-crafted audio features used. Given an impassioned speech, the generated gestures with WavLM display greater amplitude and heightened intensity, because WavLM contains rich high-level information such as emotions and semantics [37]. The final three columns of Fig. E1 show that textures are restored after refinement, especially in hands and the face.

Please refer to our homepage for more visualization results of comparison with other methods and the ablation study.

F. Capability of Generating Long Gesture Videos

To better assess the effectiveness of the optimal motion selection module and the capability of our framework to generate long gesture videos, we conduct another user study following [14]. We sample 10 long audios from the original PATS dataset as conditions to generate videos of 28s, and compare the generated results of 1) our complete framework, 2) long noise sampling (LN Samp.), 3) direct concatenation (Concat.), 4) ANGIE, and 5) the ground truth. 20 participants are asked to evaluate the effective duration of the videos, *i.e.* to decide how many seconds of the videos are effective. The average effective duration for each method is shown in Tab. F2. The results show that, although based on an easy-to-make hand-crafted rule, the optimal motion selection module benefits our method to generate longer videos with better coherency and consistency compared to only seed motion used and other methods. Directly sampling long noise and the autoregressive generation approach of ANGIE both face challenges in generating effective videos over 10 seconds.

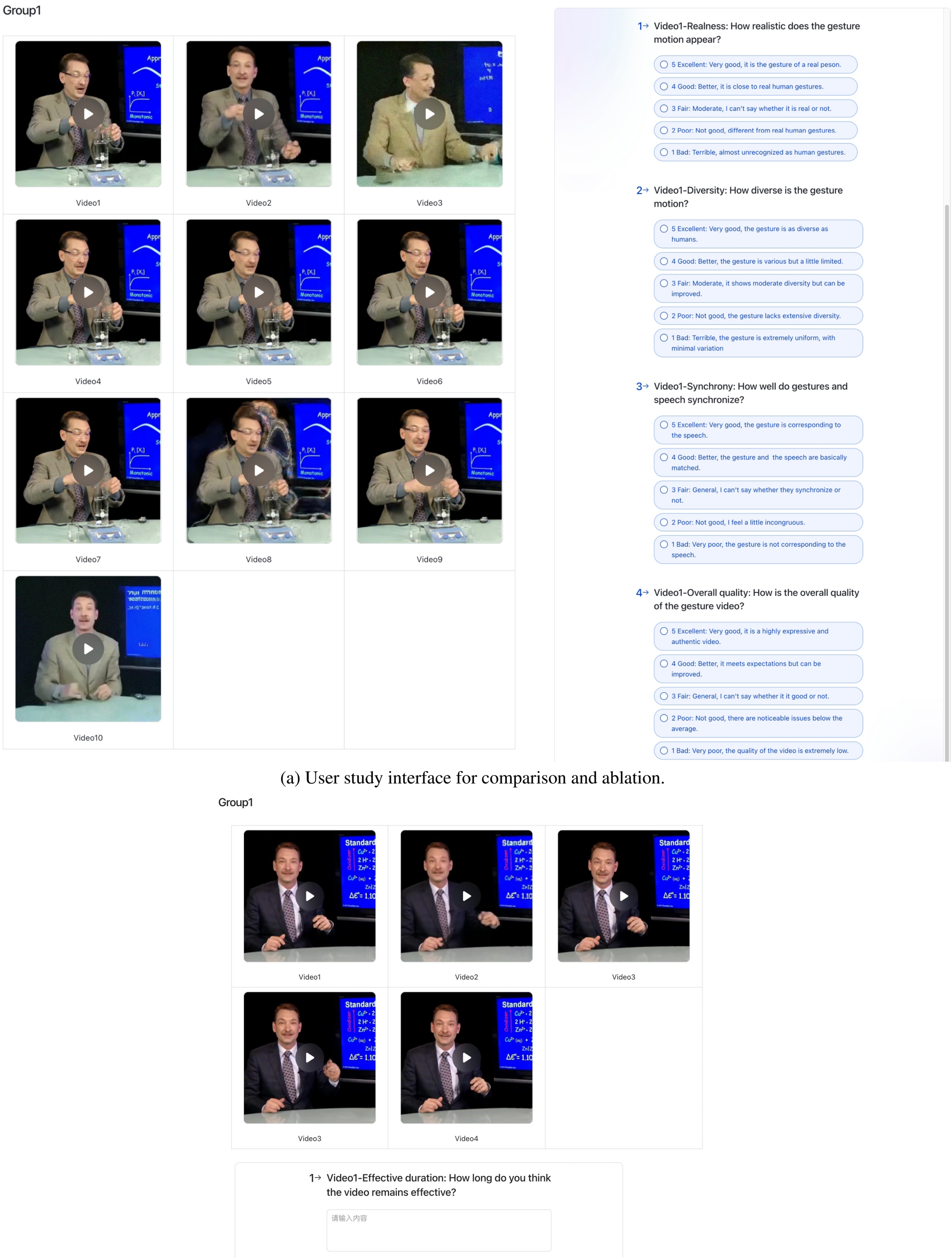
G. Details of User Study

The user study is conducted by 20 participants with good English proficiency, involving 15 males and 5 females. Each participant is remunerated about 15 USD for a rating of 40-50 minutes, which is approximately at the average wage level [39]. Screenshots of the rating interface used for comparison, the ablation study, and the evaluation of long video generation are presented in Fig. G2.

H. Robustness and Effectiveness of Objective Metrics

From the main paper, we observe: 1) ANGIE [18] achieves higher BAS than ours. 2) Refinement brings lower BAS. 3) Sampling long noise and concatenation strategies have similar BAS. All these observations regarding BAS are inconsistent with subjective perceptions. Actually, BAS considers the distance between each audio beat with its nearest gesture beat, while gesture beats are defined as local velocity minima of 2D pose sequences filtered with a Gaussian kernel [13]. In practice, we encounter unavoidable inter-frame jitters when extracting 2D poses for evaluation with the off-the-shelf pose estimator. Tremors such as those in ANGIE, blurred images without refinement, or almost stationary long noise sampling results could amplify the jitters of estimated poses and cause incorrect identification as denser gesture beats, reducing the distance between gesture and speech beats and thus incorrectly increasing BAS, which can be seen from Fig. H3. In summary, BAS is susceptible to unrelated factors, making it a less robust objective metric. FGD, Diversity, and FVD are calculated in the feature space, making them somewhat more robust compared to BAS.

Another interesting finding is that despite other metrics of our method being closer to the GT, FGD still exhibits a noticeable discrepancy. However, user study results strongly indicate the authenticity of our generated motion. One plausible explanation is that for FGD, we take the entire data, including the training and testing sets, as the real reference to calculate distribution distances. Given the rich diversity of gestures, there are inherent distribution gaps between the training and test-



1→ Video1-Realness: How realistic does the gesture motion appear?

- ○ 5 Excellent: Very good, it is the gesture of a real peson.
- ○ 4 Good: Better, it is close to real human gestures.
- ○ 3 Fair: Moderate, I can't say whether it is real or not.
- ○ 2 Poor: Not good, different from real human gestures.
- ○ 1 Bad: Terrible, almost unrecognized as human gestures.

2→ Video1-Diversity: How diverse is the gesture motion?

- ○ 5 Excellent: Very good, the gesture is as diverse as humans.
- ○ 4 Good: Better, the gesture is various but a little limited.
- ○ 3 Fair: Moderate, it shows moderate diversity but can be improved.
- ○ 2 Poor: Not good, the gesture lacks extensive diversity.
- ○ 1 Bad: Terrible, the gesture is extremely uniform, with minimal variation

3→ Video1-Synchrony: How well do gestures and speech synchronize?

- ○ 5 Excellent: Very good, the gesture is corresponding to the speech.
- ○ 4 Good: Better, the gesture and the speech are basically matched.
- ○ 3 Fair: General, I can't say whether they synchronize or not.
- ○ 2 Poor: Not good, I feel a little incongruous.
- ○ 1 Bad: Very poor, the gesture is not corresponding to the speech.

4→ Video1-Overall quality: How is the overall quality of the gesture video?

- ○ 5 Excellent: Very good, it is a highly expressive and authentic video.
- ○ 4 Good: Better, it meets expectations but can be improved.
- ○ 3 Fair: General, I can't say whether it it good or not.
- ○ 2 Poor: Not good, there are noticeable issues below the average.
- ○ 1 Bad: Very poor, the quality of the video is extremely low.

(a) User study interface for comparison and ablation.

1→ Video1-Effective duration: How long do you think the video remains effective?

请输入内容

(b) User study interface for rating effective duration.

Figure G2. Screenshots of the user study interface.

ing sets. Our model learns the data distribution from the training set, slightly deviating from the entire, while the GT of the testing set constitutes a portion of the overall distribution. This results in a noticeable difference in FGD. Referring to the training distribution reduces the difference (GT: 8.976 to 10.327 *vs*. ours: 18.131 to 13.285), providing supporting evidence.

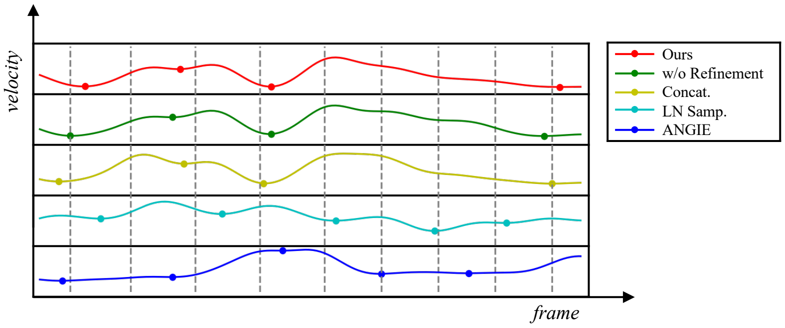


Figure H3. **Examples of velocity-frame curves of motion sequences** generated by each method for BAS analysis. For clear visualization, velocity is normalized and displayed without overlap. Dots represent gesture beats and dashed lines signify speech beats. "Concat." is short for concatenation. "LN Samp." is short for long noise sampling. "LN Samp." and "ANGIE" exhibit more gesture beats but are not aligned with speech beats.

Actually, previous studies [11, 12] indicate that co-speech gesture generation still lacks objective metrics perfectly consistent with human subjective perception. To summarize the above, we have to demonstrate that subjective evaluation remains the gold standard for co-speech gesture video generation just like any other technology in the field of human-machine interaction [11].

I. Generalization Ability

Gestures vary greatly between different speakers, so previous work typically trains an independent model for each person to capture individual styles. In contrast, we train a unified model jointly with the four speakers to ensure the scalability of our method. Experimental results indicate that even in this more challenging setting, our approach still generates gestures matching individual styles. Besides, we notice joint training brings about generalization ability to the speech of unseen speakers, which can be seen on our homepage. However, it is still hard to generalize to any given portrait at present. Yet, given two critical facts: 1) our method can animate unseen dressing appearances of the four given speakers, for the dataset contains various appearances of the same speaker, and 2) efforts like [9] on extensive multi-person datasets show stronger generalization ability to unseen portraits, we believe that our approach exhibits generalization potential, and a high-quality multi-speaker gesture video dataset may help to enhance it, which will be explored in our future work.

J. Time and Resource Consumption

Tab. J3 indicates that our training and inference time are comparable to ANGIE [18] and significantly shorter than MM-Diffusion [27]. Therefore, to the best of our knowledge, we achieve an optimal trade-off between time consumption and generation quality with distinct superiority in the latter. Although motion decoupling takes longer time, it greatly reduces the overall time and resource commitment compared to MM-Diffusion and other video generation works, *e.g*. [9] taking 14 days on 4 NVIDIA A100 GPUs for training[1], providing a relatively efficient solution. Notably, our proposed diffusion model in the latent motion space achieves competitive generation results with relatively less time consumption, highlighting its necessity in the audio-to-motion process. Undeniably, repetitive diffusion denoising introduces extra inference time, and we will further explore methods like LCM [20] and Flow Matching [16] for acceleration.

Table J3. **Time consumption comparison** of training (6 NVIDIA A10 GPUs) and inference (1 NVIDIA GeForce RTX 4090 GPUs).

Name	Training	Training Breakdown	Inference (Generate a video of ∼10 sec)
ANGIE	∼5d	Motion Representation ∼3d + Quantization ∼0.2d + Gesture GPT ∼1.8d	∼30 sec
MM-Diffusion	∼14d	Generation ∼9d + Super-Resolution ∼5d	∼600 sec
Ours	∼5d	Motion Decoupling ∼3d + Motion Diffusion ∼1.5d + Refinement ∼0.5d	∼35 sec

[1] Experimental results from our reproduced code instead of official resources.

K. Limitations and Future Work

As research towards a relatively unexplored problem, there is still room for improvements in the following areas.

Despite significant superiority to existing methods, our generated videos still exhibit some accuracy issues of blurs and flickering, especially in hand details. This arises from the intricate structures of hands, characterized by varying movements like intersections and overlaps, which actually presents an unresolved challenge in the field of image and video generation [9, 25]. TPS-based motion decoupling effectively captures curved hand contours, making our method more adaptable to complex hand shapes than ANGIE [18], but still struggles to model structural details. The limited presence of hands in the frame drawing insufficient attention, coupled with the relatively weak inpainting capability of the image synthesis network, also leads to inaccurate hands. In addition, we observe that PATS dataset sourced from in-the-wild videos is of limited quality with noticeable hand motion blur, influencing the network's performance to some extent. Therefore, in our future work, we will: 1) refine our method, *e.g.* prioritizing attention to hands and inpainting occlusion with more powerful pre-trained image generation models like SD model [25], and 2) collect high-quality gesture video data with clearer representations of hands to further enhance the generation quality.

Our current solution is unable to effectively synthesize the lip shape because there is a gap in the relationship between lips and gestures with speech. A unified framework for generating co-speech gestures and the lip shape simultaneously remains a valuable research problem, which we will explore in future work. In some showcases of the supplementary video, we use the off-the-shelf Wav2Lip [21] to synthesize lip shapes. Note that, the lip shape is not within the scope of this work, and generating lip shapes is just for better visual effects in the demo video.

For videos of bad quality, the accuracy of 2D poses from the pose estimator is compromised, leading to significant uncertainty when calculating all objective metrics regarding motion, especially BAS. Up until now, human subjective evaluation remains the most effective means of assessing generated gesture videos. Further exploration is needed to develop more robust and effective objective metrics.

L. Dataset License

We download the YouTube videos and perform preprocessing according to the video links in the metadata provided by the PATS dataset [1, 2, 6]. Video license "CC BY - NC - ND4.0 International" allows for non-commercial use. Although the video data includes personal identity information, we adhere to the data usage license, and our processed data, models, and results will be used only for academic purposes and not be permitted for commercial use.

References

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 10

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. 2020. 10

[3] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 1

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 7

[5] Philippe G Ciarlet and Pierre-Arnaud Raviart. A mixed finite element method for the biharmonic equation. In *Mathematical aspects of finite elements in partial differential equations*, pages 125–145. Elsevier, 1974. 1

[6] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 10

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[9] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 9, 10

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[11] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21, 2021. 9

[12] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 792–801, 2023. 9

[13] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2(3), 2021. 5, 7

[14] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023. 4, 7

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[16] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 9

[17] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 4

[18] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems*, 35:21386–21399, 2022. 4, 5, 6, 7, 9, 10

[19] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 5

[20] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 9

[21] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 10

[22] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086, 2021. 5

[23] Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. 3, 4

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 10

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[27] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 5, 9

[28] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20(17):10032–10044, 2020. 5

[29] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 2

[30] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 2, 6

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[32] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 3

[33] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[34] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[35] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8858–8867, 2019. 4

[36] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022. 4

[37] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. 7

[38] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 5

[39] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022. 7

[40] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3

[41] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 3

[42] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 2, 4

[43] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3418–3428, 2022. 5